Sample-Efficient Reinforcement Learning for Linearly-Parameterized MDPs with a Generative Model

Bingyan Wang*
Princeton University
bingyanw@princeton.edu

Yuling Yan*
Princeton University
yulingy@princeton.edu

Jianqing Fan
Princeton University
jqfan@princeton.edu

Abstract

The curse of dimensionality is a widely known issue in reinforcement learning (RL). In the tabular setting where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite, to obtain a nearly optimal policy with sampling access to a generative model, the minimax-optimal sample complexity scales linearly with $|\mathcal{S}| \times |\mathcal{A}|$, which can be prohibitively large when $\mathcal{S}$ or $\mathcal{A}$ is large. This paper considers a Markov decision process (MDP) that admits a set of state-action features, which can linearly express (or approximate) its probability transition kernel. We show that a model-based approach (resp. Q-learning) provably learns an ε-optimal policy (resp. Q-function) with high probability as soon as the sample size exceeds the order of $\frac{K}{(1-\gamma)^3\varepsilon^2}$ (resp. $\frac{K}{(1-\gamma)^4\varepsilon^2}$), up to some logarithmic factor. Here K is the feature dimension and $\gamma \in (0,1)$ is the discount factor of the MDP. The results is applicable to the tabular MDPs by taking the coordinate basis with $K = |\mathcal{S}| \times |\mathcal{A}|$. Both sample complexity bounds are provably tight, and our result for the model-based approach matches the minimax lower bound. Our results show that for arbitrarily large-scale MDP, both the model-based approach and Q-learning are sample-efficient when K is relatively small, and hence the title of this paper.

1 Introduction

Reinforcement learning (RL) studies the problem of learning and decision making in a Markov decision process (MDP). Recent years have seen exciting progress in applications of RL in real world decision-making problems such as AlphaGo [46, 47] and autonomous driving [33]. Specifically, the goal of RL is to search for an optimal policy that maximizes the cumulative reward, based on sequential noisy data. There are two popular approaches to RL: model-based and model-free ones.

- The model-based approaches start with formulating an empirical MDP by learning the probability transition model from the collected data samples, and then estimating the optimal policy / value function based on the empirical MDP.
- The model-free approaches (e.g. Q-learning) learn the optimal policy or the optimal (action-)value function from samples. As its name suggests, model-free approaches do not attempt to learn the model explicitly.

*Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Generally speaking, model-based approaches enjoy great flexibility since after the transition model is learned in the first place, it can then be applied to any other problems without touching the raw data samples. In comparison, model-free methods, due to its online nature, are usually memory-efficient and can interact with the environment and update the estimate on the fly.

This paper is devoted to investigating the sample efficiency of both model-based RL and Q-learning (arguably one of the most commonly adopted model-free RL algorithms). It is well known that MDPs suffer from the curse of dimensionality. For example, in the tabular setting where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite, to obtain a near optimal policy or value function given sampling access to a generative model, the minimax optimal sample complexity scales linearly with $|\mathcal{S}| \times |\mathcal{A}|$ [5, 2]. However contemporary applications of RL often encounters environments with exceedingly large state and action spaces, whilst the data collection might be expensive or even high-stake. This suggests a large gap between the theoretical findings and practical decision-making problems where $|\mathcal{S}|$ and $|\mathcal{A}|$ are large or even infinite.

To close the aforementioned theory-practice gap, one natural idea is to impose certain structural assumption on the MDP. In this paper we follow the feature-based linear transition model studied in [63], where each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ admits a K dimensional feature vector $\phi(s, a) \in \mathbb{R}^K$ that expresses the transition dynamics $\mathbb{P}(\cdot|s, a) = \Psi\phi(s, a)$ for some unknown matrix $\Psi \in \mathbb{R}^{|\mathcal{S}|\times K}$ which is common for all (s, a). This model encompasses both the tabular case and the homogeneous model in which the state space can be partitioned into K equivalent classes. Assuming access to a generative model [31, 32], under this structural assumption, this paper aims to answer the following two questions:

> *How many samples are needed for model-based RL and Q-learning to learn an optimal policy under the feature-based linear transition model?*

In what follows, we will show that the answer to this question scales linearly with the dimension of the feature space K and is independent of $|\mathcal{S}|$ and $|\mathcal{A}|$ under the feature-based linear transition model. With the aid of this structural assumption, model-based RL and Q-learning becomes significantly more sample-efficient than that in the tabular setting.

Our contributions. We focus our attention on an infinite horizon MDP with discount factor $\gamma \in (0, 1)$. We use ε-optimal policy to indicate the policy whose expected discounted cumulative rewards are ε close to the optimal value of the MDP. Our contributions are two-fold:

- We demonstrate that model-based RL provably learns an ε-optimal policy by performing planning based on an empirical MDP constructed from a total number of

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3\varepsilon^2}\right)$$

  samples, for all $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Here $\widetilde{O}(\cdot)$ hides logarithmic factors compared to the usual $O(\cdot)$ notation. To the best of our knowledge, this is the first theoretical guarantee for model-based RL under the feature-based linear transition model. This sample complexity bound matches the minimax limit established in [63] up to logarithmic factor.

- We also show that Q-learning provably finds an entrywise ε-optimal Q-function using a total number of

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^4\varepsilon^2}\right)$$

  samples, for all $\varepsilon \in (0, 1]$. This sample complexity upper bound improves the state-of-the-art result in [63] and the dependency on the effective horizon $(1-\gamma)^{-4}$ is sharp in view of [34].

These results taken collectively show the minimax optimality of model-based RL and the sub-optimality of Q-learning in sample complexity.

2 Problem formulation

This paper focuses on tabular MDPs in the discounted infinite-horizon setting [10]. Here and throughout, $\Delta_{d-1} \coloneqq \{v \in \mathbb{R}^d : \sum_{i=1}^d v_i = 1, v_i \geq 0, \forall i \in [d]\}$ stands for the d-dimensional probability simplex and $[N] \coloneqq \{1, 2, \cdots, N\}$ for any $N \in \mathbb{N}^+$.

Discounted infinite-horizon MDPs. Denote a discounted infinite-horizon MDP by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$ is a finite set of states, $\mathcal{A} = \{1, \cdots, |\mathcal{A}|\}$ is a finite set of actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{|\mathcal{S}|-1}$ represents the probability transition kernel where $P(s'|s, a)$ denotes the probability of transiting from state s to state s' when action a is taken, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the reward function where $r(s, a)$ is the instantaneous reward received when taking action $a \in \mathcal{A}$ while in state $s \in \mathcal{S}$, and $\gamma \in (0, 1)$ is the discount factor.

Value function and Q-function. Recall that the goal of RL is to learn a policy that maximizes the cumulative reward, which corresponds to value functions or Q-functions in the corresponding MDP. For a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ and a starting state $s \in \mathcal{S}$, we define the value function as

$$V^\pi(s) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \;\middle|\; s_0 = s\right]$$

for all $s \in \mathcal{S}$. Here, the trajectory is generated by $a_k = \pi(s_k)$ and $s_{k+1} \sim P(s_{k+1}|s_k, a_k)$ for every $k \geq 0$. This function measures the expected discounted cumulative reward received on the trajectory $\{(s_k, a_k)\}_{k\geq 0}$ and the expectation is taken with respect to the randomness of the transitions $s_{k+1} \sim P(\cdot|s_k, a_k)$ on the trajectory. Recall that the immediate rewards lie in $[0, 1]$, it is easy to derive that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$ for any policy π and state s. Accordingly, we define the Q-function for policy π as

$$Q^\pi(s, a) \coloneqq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \,|s_0 = s, a_0 = a\right]$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, the actions are chosen by the policy π except for the initial state (i.e. $a_k = \pi(s_k)$ for all $k \geq 1$). Similar to the value function, we can easily check that $0 \leq Q^\pi(s, a) \leq \frac{1}{1-\gamma}$ for any π and (s, a). To maximize the value function or Q function, previous literature [9, 51] establishes that there exists an optimal policy $\pi^\star$ which simultaneously maximizes $V^\pi(s)$ (resp. $Q^\pi(s, a)$) for all $s \in \mathcal{S}$ (resp. $(s, a) \in \mathcal{S} \times \mathcal{A}$). We define the optimal value function $V^\star$ and optimal Q-function $Q^\star$ respectively as

$$V^\star(s) \coloneqq \max_\pi V^\pi(s) = V^{\pi^\star}(s), \qquad Q^\star(s, a) \coloneqq \max_\pi Q^\pi(s, a) = Q^{\pi^\star}(s, a)$$

for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Linear transition model. Given a set of K feature functions $\phi_1, \phi_2, \cdots, \phi_K$: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define ϕ to be a feature mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^K$ such that

$$\phi(s, a) = [\phi_1(s, a), \cdots, \phi_K(s, a)] \in \mathbb{R}^K.$$

Then we are ready to define the linear transition model [63] as follows.

Definition 1 (Linear transition model)**.** *Given a discounted infinite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$, M admits the linear transition model if there exists some (unknown) functions $\psi_1, \cdots, \psi_K : \mathcal{S} \to \mathbb{R}$, such that*

$$P(s'|s, a) = \sum_{k=1}^{K} \phi_k(s, a)\, \psi_k(s') \tag{1}$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$.

Readers familiar with linear MDP literatures might immediately recognize that the above definition is the same as the structure imposed on the probability transition kernel P in the linear MDP model [63, 30, 65, 26, 52, 56, 60]. However unlike linear MDP which also requires the reward function $r(s, a)$ to be linear in the feature mapping $\phi(s, a)$, here we do not impose any structural assumption on the reward.

Example 1 (Tabular MDP). *Each tabular MDP can be viewed as a linear transition model with feature mapping $\phi(s,a) = \boldsymbol{e}_{(s,a)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ (i.e. the vector with all entries equal to 0 but the one corresponding to (s,a) equals to 1) for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. To see this, we can check that Definition 1 is satisfied with $K = |\mathcal{S}|\times|\mathcal{A}|$ and $\psi_{(s,a)}(s') = \mathbb{P}(s'|s,a)$ for each $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. This example is a sanity check of Definition 1, which also shows that our results (Theorem 1 and 2) can recover previous results on tabular MDP [2, 34] by taking $K = |\mathcal{S}|\times|\mathcal{A}|$.*

Example 2 (Simplex Feature Space). *If all feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ fall in the probability simplex Δ_{K-1}, a linear transition model can be constructed by taking $\psi_k(\cdot)$ to be any probability measure over $\mathcal{S}$ for all $k \in [K]$.*

A key observation is that the model size of linear transition model with known feature mapping ϕ is $|\mathcal{S}|K$ (the number of coefficients $\psi_k(s')$ in (1)), which is still large when the state space $\mathcal{S}$ is large. In contrast, it will be established later that to learn a near-optimal policy or Q-function, we only need a much smaller number of samples, which depends linearly on K and is independent of $|\mathcal{S}|$.

Next, we introduce a critical assumption employed in prior literature [63, 67, 45].

Assumption 1 (Anchor state-action pairs). *Assume there exists a set of anchor state-action pairs $\mathcal{K} \subset \mathcal{S}\times\mathcal{A}$ with $|\mathcal{K}| = K$[2] such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, its corresponding feature vector can be expressed as a convex combination of the feature vectors of anchor state-action pairs $\{(s,a) : (s,a) \in \mathcal{K}\}$:*

$$\phi(s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i(s,a)\,\phi(s_i,a_i) \quad \textit{for} \quad \sum_{i=1}^{K} \lambda_i(s,a) = 1 \quad \textit{and} \quad \lambda_i(s,a) \geq 0. \tag{2}$$

Further, we assume that the vectors in $\{\phi(s,a) : (s,a) \in \mathcal{K}\}$ are linearly independent.

We pause to develop some intuition of this assumption using Examples 1 and 2. In Example 1, it is straightforward to check that tabular MDPs satisfies Assumption 1 with $\mathcal{K} = \mathcal{S}\times\mathcal{A}$. In terms of Example 2, without loss of generality we can assume that the subspace spanned by the features has full rank, i.e. $\mathsf{span}\{\phi(s,a) : (s,a) \in \mathcal{S}\times\mathcal{A}\} = \mathbb{R}^K$ (otherwise we can reduce the dimension of feature space). Then we can also check that Example 2 satisfies Assumption 1 with arbitrary $\mathcal{K} \subseteq \mathcal{S}\times\mathcal{A}$ such that the vectors in $\{\phi(s,a) : (s,a) \in \mathcal{K}\}$ are linearly independent. In fact, this sort of "anchor" notion appears widely in the literature: [3] considers "anchor word" in topic modeling; [19] defines "separability" in their study of non-negative matrix factorization; [48] introduces "aggregate" in reinforcement learning; [21] studies "anchor state" in soft state aggregation models. These concepts all bear some kind of resemblance to our definition of anchor state-action pairs here.

Throughout this paper, we assume that the feature mapping ϕ is known, which is a widely adopted assumption in previous literature [63, 30, 68, 26, 52, 56, 60]. In practice, large scale RL usually makes use of representation learning to obtain the feature mapping ϕ. Furthermore, the learned representations can be selected to satisfy the anchor state-action pairs assumption by design.

A useful implication of Assumption 1 is that we can represent the transition kernel as

$$P(\cdot|s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i(s,a)\,P(\cdot|s_i,a_i), \tag{3}$$

This follows simply from substituting (2) into (1) (see (14) in Appendix A for a formal proof).

3 Model-based RL with a generative model

We start with studying model-based RL with a generative model in this section. We propose a model-based planning algorithm and show that it returns an ε-optimal policy with minimax optimal sample size.

3.1 Main results

A generative model and an empirical MDP. We assume access to a generative model that provides us with independent samples from M. For each anchor state-action pair $(s_i,a_i) \in \mathcal{K}$, we collect N

[2] Without loss of generality, one can always assume that the number of anchor state-action pairs equals to the feature dimension K. Interested readers are referred to Appendix D for detailed argument.

independent samples $s_i^{(j)} \sim P(\cdot|s_i, a_i)$, $j \in [N]$. This allows us to construct an empirical transition kernel $\widehat{P}$ where

$$\widehat{P}(s' \mid s, a) = \sum_{i=1}^{K} \lambda_i(s, a) \cdot \left(\frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\left\{ s_i^{(j)} = s' \right\} \right), \tag{4}$$

for each $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, $\frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\{s_i^{(j)} = s'\}$ is an empirical estimate of $P(s'|s_i, a_i)$ and then (3) is employed. With $\widehat{P}$ in hand, we can construct an empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$. Our goal here is to derive the sample complexity which guarantees that the optimal policy of $\widehat{M}$ is an ε-optimal policy for the true MDP M. The algorithm is summarized below.

Algorithm 1 Model-based RL with a generative model

Inputs: a set of anchor state-action pairs $\mathcal{K} = \{(s_i, a_i) : i \in [K]\}$, feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$, any planning algorithm $\mathcal{P}$, target algorithmic error level $\varepsilon_{\mathsf{opt}}$.
For $i = 1, \cdots, K$ **do**
  Draw N independent samples $s_i^{(j)} \sim P(\cdot|s_i, a_i)$, $j = 1, \cdots, N$.
End for
Construct an empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$ where $\widehat{P}$ can be computed by (4).
Output $\widehat{\pi}$ as an $\varepsilon_{\mathsf{opt}}$-optimal policy of $\widehat{M}$ computed by the planning algorithm $\mathcal{P}$.

Careful readers may note that in Algorithm 1, $\{\lambda(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ is used in the construction of $\widehat{P}$, while $\{\lambda(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ is not input into the algorithm. This is because given $\mathcal{K}$ and ϕ are known, $\{\lambda(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ can be calculated explicitly. The following theorem provides theoretical guarantees for the output policy $\widehat{\pi}$ of the chosen optimization algorithm on the empirical MDP $\widehat{M}$.

Theorem 1. *Suppose that $\delta > 0$ and $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Let $\widehat{\pi}$ be the policy returned by Algorithm 1. Assume that*

$$N \geq \frac{C \log\left(K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)^3 \varepsilon^2} \tag{5}$$

for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - \delta$,

$$Q^\star(s, a) - Q^{\widehat{\pi}}(s, a) \leq \varepsilon + \frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma}, \tag{6}$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here $\varepsilon_{\mathsf{opt}}$ is the target algorithmic error level in Algorithm 1.

We first remark that the two terms on the right hand side of (6) can be viewed as statistical error and algorithmic error, respectively. The first term ε denotes the statistical error coming from the deviation of the empirical MDP $\widehat{M}$ from the true MDP M. As the sample size N grows, ε could decrease towards 0. The other term $4\varepsilon_{\mathsf{opt}}/(1-\gamma)$ represents the algorithmic error where $\varepsilon_{\mathsf{opt}}$ is the target accuracy level of the planning algorithm applied to $\widehat{M}$. Note that $\varepsilon_{\mathsf{opt}}$ can be arbitrarily small if we run the planning algorithm (e.g. value iteration) for enough iterations. A few implications of this theorem are in order.

- *Minimax-optimal sample complexity*. Assume that $\varepsilon_{\mathsf{opt}}$ is made negligibly small, e.g. $\varepsilon_{\mathsf{opt}} = O((1-\gamma)\varepsilon)$ to be discussed in the next point. Note that we draw N independent samples for each state-action pair $(s, a) \in \mathcal{K}$, therefore the requirement (5) for finding an $O(\varepsilon)$-optimal policy translates into the following sample complexity requirement

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3 \varepsilon^2} \right).$$

  This matches the minimax optimal lower bound (up to a logarithm factor) established in [63, Theorem 1] for feature-based MDP. In comparison, for tabular MDP the minimax optimal sample complexity is $\widetilde{\Omega}((1-\gamma)^{-3}\varepsilon^{-2}|\mathcal{S}||\mathcal{A}|)$ [5, 2]. Our sample complexity scales linearly with K instead of $|\mathcal{S}||\mathcal{A}|$ for tabular MDP as desired.

- *Computational complexity*. An advantage of Theorem 1 is that it incorporates the use of any efficient planning algorithm applied to the empirical MDP $\widehat{M}$. Classical algorithms include Q-value iteration (QVI) or policy iteration (PI) [43]. For example, QVI achieves the target level $\varepsilon_{\mathsf{opt}}$ in $O((1-\gamma)^{-1}\log\varepsilon_{\mathsf{opt}}^{-1})$ iterations, and each iteration takes time proportional to $O(NK+|\mathcal{S}||\mathcal{A}|K)$. To learn an $O(\varepsilon)$-optimal policy, which requires sample complexity (5) and the target level $\varepsilon_{\mathsf{opt}} = O((1-\gamma)\varepsilon)$, the overall running time is

$$\widetilde{O}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|\,K}{1-\gamma}+\frac{K}{(1-\gamma)^4\,\varepsilon^2}\right).$$

  In comparison, for the tabular MDP the corresponding running time is $\widetilde{O}((1-\gamma)^{-4}\varepsilon^{-2}|\mathcal{S}||\mathcal{A}|)$ [2]. This suggests that under the feature-based linear transition model, the computational complexity is $\min\{|\mathcal{S}||\mathcal{A}|/K,(1-\gamma)^{-3}\varepsilon^{-2}/K\}$ times lower than that for the tabular MDP (up to logarithm factors), which is significantly more efficient when K is not too large.
- *Stability vis-à-vis model misspecification*. A more general version of Theorem 1 (Theorem 3 in Appendix B) shows that when P approximately (instead of exactly) admits the linear transition model, we can still achieve some meaningful result. Specifically, if there exists a linear transition kernel $\widetilde{P}$ obeying $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\widetilde{P}(\cdot|s,a)-P(\cdot|s,a)\|_1\le\xi$ for some $\xi\ge 0$, we can show that $\widehat{\pi}$ returned by Algorithm 1 (with slight modification) satisfies

$$Q^\star(s,a)-Q^{\widehat{\pi}}(s,a)\le\varepsilon+\frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma}+\frac{22\xi}{(1-\gamma)^2},$$

  for every $(s,a)\in\mathcal{S}\times\mathcal{A}$. This shows that the model-based method is stable vis-a-vis model misspecification. Interested readers are referred to Appendix B for more details.

In Algorithm 1, the reward function r is assumed to be known. If the information of r is unavailable, an alternative is to assume that r is linear with respect to the feature mapping ϕ, i.e. $r(s,a)=\theta^\top\phi(s,a)$ for every $(s,a)\in\mathcal{S}\times\mathcal{A}$, which is widely adopted in linear MDP literature [26, 30, 56, 60]. Under this linear assumption, one can obtain θ by solving the following linear system of equations

$$r(s,a)=\theta^\top\phi(s,a),\quad \forall(s,a)\in\mathcal{K}, \tag{7}$$

which can be constructed by the observed reward $r(s,a)$ for all anchor state-action pairs.

4 Model-free RL—vanilla Q Learning

In this section, we turn to study one of the most popular model-free RL algorithms—Q-learning. We provide tight sample complexity bound for vanilla Q-learning under the feature-based linear transition model, which shows its sample-efficiency (depends on $|K|$ instead of $|\mathcal{S}|$ or $|\mathcal{A}|$) and sub-optimality in the dependency on the effective horizon.

4.1 Q-learning algorithm

The vanilla Q-learning algorithm maintains a Q-function estimate $Q_t:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ for all $t\ge 0$, with initialization Q_0 obeying $0\le Q_0(s,a)\le\frac{1}{1-\gamma}$ for every $(s,a)\in\mathcal{S}\times\mathcal{A}$. Assume we have access to a generative model. In each iteration $t\ge 1$, we collect an independent sample $s_t(s,a)\sim P(\cdot|s,a)$ for every anchor state-action pair $(s,a)\in\mathcal{K}$ and define $Q_{\mathcal{K}}^{(t)}:\mathcal{K}\to\mathbb{R}$ to be

$$Q_{\mathcal{K}}^{(t)}(s,a)\coloneqq\max_{a'\in\mathcal{A}}Q_t(s_t,a'),\qquad s_t\equiv s_t(s,a)\sim P(\cdot|s,a).$$

Then given the learning rate $\eta_t\in(0,1]$, the algorithm adopts the following update rule to update all entries of the Q-function estimate

$$Q_t=(1-\eta_t)\,Q_{t-1}+\eta_t\mathcal{T}_{\mathcal{K}}^{(t)}(Q_{t-1}).$$

Here, $\mathcal{T}_{\mathcal{K}}^{(t)}$ is an empirical Bellman operator associated with the linear transition model M and the set $\mathcal{K}$ and is given by

$$\mathcal{T}_{\mathcal{K}}^{(t)}(Q)(s,a)\coloneqq r(s,a)+\gamma\lambda(s,a)\,Q_{\mathcal{K}}^{(t)},$$

where (3) is used in the construction. Clearly, this newly defined operator $\mathcal{T}_{\mathcal{K}}^{(t)}$ is an unbiased estimate of the famous Bellman operator $\mathcal{T}$ [8] defined as

$$\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}:\qquad \mathcal{T}\,(Q)\,(s,a)\coloneqq r\,(s,a)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}}Q\,(s',a')\right].$$

A critical property is that the Bellman operator $\mathcal{T}$ is contractive with a unique fixed point which is the optimal Q-function $Q^{\star}$[8]. To solve the fixed-point equation $\mathcal{T}(Q^{\star})=Q^{\star}$, Q-learning was then introduced by [58] based on the idea of stochastic approximation [44]. This procedure is precisely described in Algorithm 2.

Algorithm 2 Vanilla Q-learning for infinite-horizon discounted MDPs

inputs: learning rates $\{\eta_t\}$, number of iterations T, discount factor γ, initial estimate Q_0.
for $t=1,\ldots,T$ **do**
  Draw $s_t\,(s,a)\sim P\,(\cdot|s,a)$ for each $(s,a)\in\mathcal{K}$.
  Compute $\boldsymbol{Q}_t$ according to the update rule

$$Q_t=(1-\eta_t)\,Q_{t-1}+\eta_t\mathcal{T}_{\mathcal{K}}^{(t)}\,(Q_{t-1})\,.$$

end for

4.2 Main results

We are now ready to provide our main result for vanilla Q-learning, assuming sampling access to a generative model.

Theorem 2. *Consider any $\delta\in(0,1)$ and $\varepsilon\in(0,1]$. Assume that for any $0\le t\le T$, the learning rates satisfy*

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}}\le\eta_t\le\frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}} \tag{8}$$

for some sufficiently small universal constants $c_1\ge c_2>0$. Suppose that the total number of iterations T exceeds

$$T\ge\frac{C_3\log\left(KT/\delta\right)\log^4 T}{(1-\gamma)^4\,\varepsilon^2} \tag{9}$$

for some sufficiently large universal constant $C_3>0$. If the initialization obeys $0\le Q_0(s,a)\le\frac{1}{1-\gamma}$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, then with probability exceeding $1-\delta$, the output Q_T of Algorithm 2 satisfies

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}|Q_T\,(s,a)-Q^{\star}\,(s,a)|\le\varepsilon. \tag{10}$$

In addition, let π_T (resp. V_T) to be the policy (resp. value function) induced by Q_T, then one has

$$\max_{s\in\mathcal{S}}|V^{\pi_T}\,(s)-V^{\star}\,(s)|\le\frac{2\gamma\varepsilon}{1-\gamma}. \tag{11}$$

This theorem provides theoretical guarantees on the performance of Algorithm 2. A few implications of this theorem are in order.

- *Learning rate*. The condition (8) accommodates two commonly adopted choice of learning rates: (i) linearly rescaled learning rates $\eta_t=[1+c_2(1-\gamma)t/\log^2 T]^{-1}$, and (ii) iteration-invariant learning rates $\eta_t\equiv[1+c_1(1-\gamma)T/\log^2 T]$. Interested readers are referred to the discussions in [34, Section 3.1] for more details on these two learning rate schemes.
- *Tight sample complexity bound*. Note that we draw K independent samples in each iteration, therefore the iteration complexity (9) can be translated into the sample complexity bound TK in order for Q-learning to achieve ε-accuracy:

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^4\,\varepsilon^2}\right). \tag{12}$$

  As we will see shortly, this result improves the state-of-the-art sample complexity bound presented in [63, Theorem 2] . In addition, the dependency on the effective horizon $(1-\gamma)^{-4}$ matches the lower bound established in [34, Theorem 2] for vanilla Q-learning using either learning rate scheme covered in the previous remark, suggesting that our sample complexity bound (12) is sharp.

- *Stability vis-à-vis model misspecification*. Just like the model-based approach, we can also show that Q-learning is also stable vis-a-vis model misspecification when P approximately admits the linear transition model. We refer interested readers to Theorem 4 in Appendix B for more details.

Comparison with [63]. We compare our result with the sample complexity bounds for Q-learning under the feature-based linear transition model in [63].

- We first compare our result with [63, Theorem 2], which is, to the best of our knowledge, the state-of-the-art theory for this problem. When there is no model misspecification, [63, Theorem 2] showed that in order for their Phased Parametric Q-learning[3] (Algorithm 1 therein) to learn an ε-optimal policy, the sample size needs to be

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^7 \varepsilon^2}\right).$$

  Note that (12) is the sample complexity required for entrywise ε-accurate estimate of the optimal Q-function, thus a fair comparison requires to use the sample complexity for learning an ε-optimal policy deduced from (11), which is

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^6 \varepsilon^2}\right).$$

  Hence, our sample complexity improves upon previous work by a factor at least on the order of $(1-\gamma)^{-1}$. However it is worth mentioning that [63, Theorem 2] is built upon weaker conditions $\sum_{i=1}^{K} \lambda_i(s,a) = 1$ and $\sum_{i=1}^{K} |\lambda_i(s,a)| \le L$ for some $L \ge 1$, which does not require $\lambda_i(s,a) \ge 0$. Our result holds under Assumption 1, which requires $\sum_{i=1}^{K} \lambda_i(s,a) = 1$ and $\lambda_i(s,a) \ge 0$. Under the current analysis framework, it is difficult to obtain tight sample complexity bounds without assuming $\lambda_i(s,a) \ge 0$.
- Besides vanilla Q-learning, [63] also proposed a new variant of Q-learning called Optimal Phased Parametric Q-Learning (Algorithm 2 therein), which is essentially Q-learning with variance reduction. [63, Theorem 3] showed that the sample complexity for this algorithm is

$$\widetilde{O}\left(\frac{K}{(1-\gamma)^3 \varepsilon^2}\right),$$

  which matches minimax optimal lower bound (up to a logarithm factor) established in [63, Theorem 1]. Careful reader might notice that this sample complexity bound is better than ours for vanilla Q-learning. We emphasize that as elucidated in the second implication under Theorem 2, our result is already tight for vanilla Q-learning. This observation reveals that while the sample complexity for vanilla Q-learning is provably sub-optimal, the variants of Q-learning can have better performance and achieve minimax optimal sample complexity.

We conclude this section by comparing model-based and model-free approaches. Theorem 1 shows that the sample complexity of the model-based approach is minimax optimal, whilst vanilla Q-learning, perhaps the most commonly adopted model-free method, is sub-optimal according to Theorem 2. However this does not mean that model-based method is better than model-free ones since (i) some variants of Q-learning (see [63, Algorithm 2] for example) also has minimax optimal sample complexity; and (ii) in many applications it might be unrealistic to estimate the model in advance.

5 A glimpse of our technical approaches

The establishment of Theorems 1 and 2 calls for a series of technical novelties in the proof. In what follows, we briefly highlight our key technical ideas and novelties.

[3] The difference between Algorithm 2 and Phased Parametric Q-Learning in [63] is that Algorithm 2 maintains and updates a Q-function estimate Q_t, while Phased Parametric Q-Learning parameterized Q-function by

$$Q_w(s,a) := r(s,a) + \gamma \phi(s,a)^\top w,$$

and then updates the parameters w.

- For the model-based approach, we employ "leave-one-out" analysis to decouple the complicated statistical dependency between the empirical probability transition model $\widehat{P}$ and the corresponding optimal policy. Specifically, [2] proposed to construct a collection of auxiliary MDPs where each one of them leaves out a single state s by setting s to be an absorbing state and keeping everything else untouched. We tailor this high level idea to the needs of linear transition model, then the independence between the newly constructed MDP with absorbing state s and data samples collected at state s will facilitate our analysis, as detailed in Lemma 1. Compared with [2], Theorem 1 extends the tabular setting studied in [2] to the linear transition model and accommodates model misspecification, which actually needs significant efforts as detailed in the supplementary materials. This "leave-one-out" type of analysis has been utilized in studying numerous problems by a long line of work, such as [22, 38, 53, 13, 12, 14, 15], just to name a few.

- To obtain tighter sample complexity bound than the previous one $\widetilde{O}(\frac{K}{(1-\gamma)^7\varepsilon^2})$ in [63] for vanilla Q-learning, we invoke Freedman's inequality [24] for the concentration of an error term with martingale structure as illustrated in Appendix C, while the classical ones used in analyzing Q-learning are Hoeffding's inequality and Bernstein's inequality [63]. The use of Freedman's inequality helps us establish a recursive relation on $\{\|Q_t - Q^\star\|_\infty\}_{t=0}^T$, which consequently leads to the performance guarantee (10). It is worth mentioning that [34] also studied vanilla Q-learning in the tabular MDP setting and adopted Freedman's inequality, while we emphasize that it requires a lot of efforts and more delicate analyses in order to study linear transition model and also allow for model misspecification in the current paper, as detailed in the supplementary material.

6 Additional related literature

To remedy the issue of prohibitively high sample complexity, there exists a substantial body of literature proposing and studying many structural assumptions and complexity notions under different settings. This current paper focuses on linear transition model which is studied in MDP by numerous previous works [63, 30, 64, 68, 40, 25, 56, 52, 26, 60]. Among them, [63] studied linear transition model and provided tight sample complexity bounds for a new variant of Q-learning with the help of variance reduction. [30] focused on linear MDP and designed an algorithm called "Least-Squares Value Iteration with UCB" with both polynomial runtime and polynomial sample complexity without accessing generative model. [56] extended the study of linear MDP to the framework of reward-free reinforcement learning. [68] considered a different feature mapping called linear kernel MDP and devised an algorithm with polynomial regret bound without generative model. Other popular structure assumptions include: [61] studied fully deterministic transition dynamics; [28] introduced Bellman rank and proposed an algorithm which needs sample size polynomially dependent on Bellman rank to obtain a near-optimal policy in contextual decision processes; [20] assumed that the value function has low variance compared to the mean for all deterministic policy; [39, 42, 7, 66] used linear model to approximate the value function; [35] assumed that the optimal Q-function can be linearly-parameterized by the features.

Apart from the linear transition model, another notion adopted in this work is the generative model, whose role in discounted MDP has been studied by extensive literature. The concept of generative model was originally introduced by [32], and then widely adopted in numerous works, including [31, 5, 63, 54, 2, 41], to name a few. Specifically, it is assumed that a generative model of the studied MDP is available and can be queried for every state-action pair and output the next state. Among previous works, [5] proved that the minimax lower bound on the sample complexity to obtain an ε-optimal policy was $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$. [5] also showed that model-based approach can output an ε-optimal value function with near-optimal sample complexity for $\varepsilon \in (0,1)$. Then [2] made significant progress on the challenging problem of establishing minimax optimal sample complexity in estimating an ε-optimal policy with the help of "leave-one-out" analysis.

In addition, after being proposed in [59], Q-learning has become the focus of a rich line of research [58, 11, 32, 23, 4, 29, 53, 16, 37, 62]. Among them, [16, 37, 62] studied Q-learning in the presence of Markovian data, i.e. a single sample trajectory. In contrast, under the generative setting of Q-learning where a fresh sample can be drawn from the simulator at each iteration, [54] analyzed a variant of Q-learning with the help of variance reduction, which was proved to enjoy minimax optimal sample complexity $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$. Then more recently, [34] improved the lower bound of the

vanilla Q-learning algorithm in terms of its scaling with $\frac{1}{1-\gamma}$ and proved a matching upper bound $\widetilde{O}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2})$.

7 Discussion

This paper studies sample complexity of both model-based and model-free RL under a discounted infinite-horizon MDP with feature-based linear transition model. We establish tight sample complexity bounds for both model-based approaches and Q-learning, which scale linearly with the feature dimension K instead of $|\mathcal{S}| \times |\mathcal{A}|$, thus considerably reduce the required sample size for large-scale MDPs when K is relatively small. Our results are sharp, and the sample complexity bound for the model-based approach matches the minimax lower bound. The current work suggests a couple of directions for future investigation, as discussed in detail below.

- *Extension to episodic MDPs.* An interesting direction for future research is to study linear transition model in episodic MDP. This focus of this work is infinite-horizon discounted MDPs, and hopefully the analysis here can be extended to study the episodic MDP as well ([17, 18, 6, 27, 55, 26]).
- *Continuous state and action space.* The state and action spaces in this current paper are still assumed to be finite, since the proof relies heavily on the matrix operations. However, we expect that the results can be extended to accommodate continuous state and action space by employing more complicated analysis.
- *Accommodating entire range of ε.* Since both value functions and Q-functions can take value in $[0, (1-\gamma)^{-1}]$, ideally our theory should cover all choices of $\varepsilon \in (0, (1-\gamma)^{-1}]$. However we require that $\varepsilon \in (0, (1-\gamma)^{-1/2}]$ in Theorem 1 and $\varepsilon \in (0, 1]$ in Theorem 2. While most of the prior works like [2, 63] also impose these restrictions, a recent work [36] proposed a perturbed model-based planning algorithm and proved minimax optimal guarantees for any $\varepsilon \in (0, (1-\gamma)^{-1}]$. While their work only focused on model-based RL under tabular MDP, an interesting future direction is to improve our theory to accommodate any $\varepsilon \in (0, (1-\gamma)^{-1}]$.
- *General function approximation.* Another future direction is to extend the study to more general function approximation starting from linear structure covered in this paper. There exists a rich body of work proposing and studying different structures, such as linear value function approximation [39, 42, 7, 66], linear MDPs with infinite dimensional features [1], Eluder dimension [57], Bellman rank [28] and Witness rank [50], etc. Therefore, it is hopeful to investigate these settings and improve the sample efficiency.

Acknowledgments and Disclosure of Funding

B. Wang is supported in part by Gordon Y. S. Wu Fellowships in Engineering. Y. Yan is supported in part by ARO grant W911NF-20-1-0097 and NSF grant CCF-1907661. Part of this work was done while Y. Yan was visiting the Simons Institute for the Theory of Computing. J. Fan is supported in part by the ONR grant N00014-19-1-2120 and the NSF grants DMS-1662139, DMS-1712591, DMS-2052926, DMS-2053832, and the NIH grant 2R01-GM072611-15.

References

[1] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

[4] Mohammad Gheshlaghi Azar, Remi Munos, M Ghavamzadaeh, and Hilbert J Kappen. Speedy q-learning. 2011.

[5] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[7] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[8] Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

[9] Richard Bellman and Stuart Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, pages 247–251, 1959.

[10] Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

[11] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[12] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

[13] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

[14] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484, accepted to Annals of Statistics*, 2020.

[15] Yuxin Chen, Jianqing Fan, Bingyan Wang, and Yuling Yan. Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution under random designs. *Journal of the American Statistical Association*, (just-accepted):1–27, 2021.

[16] Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.

[17] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28:2818–2826, 2015.

[18] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5717–5727, 2017.

[19] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *17th Annual Conference on Neural Information Processing Systems, NIPS 2003*. Neural information processing systems foundation, 2004.

[20] Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8060–8070, 2019.

[21] Yaqi Duan, Zheng Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

[23] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for q-learning. *Journal of machine learning Research*, 5(1), 2003.

[24] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[25] Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pages 4063–4073. PMLR, 2021.

[26] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

[27] Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

[28] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[29] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

[30] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[31] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

[32] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002, 1999.

[33] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[34] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[35] Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting. *arXiv preprint arXiv:2105.08024*, 2021.

[36] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

[37] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 2020.

[38] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[39] Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

[40] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[41] Ashwin Pananjady and Martin J Wainwright. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

[42] Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759, 2008.

[43] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[44] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[45] Roshan Shariff and Csaba Szepesvári. Efficient planning in large mdps with weak linear function approximation. *arXiv preprint arXiv:2007.06184*, 2020.

[46] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[47] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[48] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems 7*, 7:361, 1995.

[49] Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

[50] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

[51] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[52] Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

[53] Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[54] Martin J Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[55] Ruosong Wang, Simon S Du, Lin Yang, and Sham Kakade. Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33, 2020.

[56] Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

[57] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

[58] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[59] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

[60] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

[61] Zheng Wen and Benjamin Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

[62] Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020.

[63] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[64] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[65] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[66] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[67] Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Limiting extrapolation in linear approximate value iteration. *Advances in Neural Information Processing Systems*, 32:5615–5624, 2019.

[68] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

Supplement to "Sample-Efficient Reinforcement Learning for Linearly-Parameterized MDPs with a Generative Model"

Bingyan Wang*
Princeton University
bingyanw@princeton.edu

Yuling Yan*
Princeton University
yulingy@princeton.edu

Jianqing Fan
Princeton University
jqfan@princeton.edu

A Notations

In this section we gather the notations that will be used throughout the appendix.

For any vectors $\boldsymbol{u} = [u_i]_{i=1}^n \in \mathbb{R}^n$ and $\boldsymbol{v} = [u_i]_{i=1}^n \in \mathbb{R}^n$, let $\boldsymbol{u} \circ \boldsymbol{v} = [u_i v_i]_{i=1}^n$ denote the Hadamard product of $\boldsymbol{u}$ and $\boldsymbol{v}$. We slightly abuse notations to use $\sqrt{\cdot}$ and $|\cdot|$ to define entry-wise operation, i.e. for any vector $\boldsymbol{v} = [v_i]_{i=1}^n$ denote $\sqrt{\boldsymbol{v}} \coloneqq [\sqrt{v_i}]_{i=1}^n$ and $|\boldsymbol{v}| \coloneqq [|v_i|]_{i=1}^n$. Furthermore, the binary notations $\leq$ and $\geq$ are both defined in entry-wise manner, i.e. $\boldsymbol{u} \leq \boldsymbol{v}$ (resp. $\boldsymbol{u} \geq \boldsymbol{v}$) means $u_i \leq v_i$ (resp. $u_i \geq v_i$) for all $1 \leq i \leq n$. For a collection of vectors $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m \in \mathbb{R}^n$ with $\boldsymbol{v}_i = [v_{i,j}]_{j=1}^n \in \mathbb{R}^n$, we define the max operator to be $\max_{1\leq i\leq m} \boldsymbol{v}_i \coloneqq [\max_{1\leq i\leq m} v_{i,j}]_{j=1}^n$.

For any matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{M}\|_1$ is defined as the largest row-wise ℓ_1 norm of $\boldsymbol{M}$, i.e. $\|\boldsymbol{M}\|_1 \coloneqq \max_i \sum_j |M_{i,j}|$. In addition, we define $\mathbf{1}$ to be a vector with all the entries being 1, and $\boldsymbol{I}$ be the identity matrix. To express the probability transition function P in matrix form, we define the matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ to be a matrix whose (s,a)-th row $\boldsymbol{P}_{s,a}$ corresponds to $P(\cdot|s,a)$. In addition, we define $\boldsymbol{P}^\pi$ to be the probability transition matrix induced by policy π, i.e. $\boldsymbol{P}^\pi_{(s,a),(s',a')} = \boldsymbol{P}_{s,a}(s')\,\mathbb{1}_{\pi(s')=a'}$ for all state-action pairs (s,a) and (s',a'). We define π_t to be the policy induced by Q_t, i.e. $Q_t(s,\pi_t(s)) = \max_a Q_t(s,a)$ for all $s \in \mathcal{S}$. Furthermore, we denote the reward function r by vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, i.e. the (s,a)-th element of $\boldsymbol{r}$ equals $r(s,a)$. In the same manner, we define $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent V^π, $V^\star$, V_t, Q^π, $Q^\star$ and Q_t respectively. By using these notations, we can rewrite the Bellman equation as

$$\boldsymbol{Q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}^\pi \boldsymbol{Q}^\pi. \tag{11}$$

Further, for any vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, let $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be

$$\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \coloneqq \boldsymbol{P}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}\boldsymbol{V})\circ(\boldsymbol{P}\boldsymbol{V}), \tag{12}$$

and define $\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}) \in \mathbb{R}$ to be

$$\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}) \coloneqq \boldsymbol{P}_{s,a}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_{s,a}\boldsymbol{V})^2, \tag{13}$$

where $\boldsymbol{P}_{s,a}$ is the (s,a)-th row of $\boldsymbol{P}$.

Next, we reconsider Assumption 1. For any state-action pair (s,a), we define vector $\boldsymbol{\lambda}(s,a) \in \mathbb{R}^K$ (resp. $\boldsymbol{\phi}(s,a) \in \mathbb{R}^K$) with $\boldsymbol{\lambda}(s,a) = [\lambda_i(s,a)]_{i=1}^K$ (resp. $\boldsymbol{\phi}(s,a) = [\phi_i(s,a)]_{i=1}^K$) and matrix

*Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

$\boldsymbol{\Lambda} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$ (resp. $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$) whose (s,a)-th row corresponds to $\boldsymbol{\lambda}(s,a)^\top$ (resp. $\boldsymbol{\phi}(s,a)^\top$). Define vector $\boldsymbol{\psi}(s,a) \in \mathbb{R}^K$ with $\boldsymbol{\psi}(s,a) = [\psi_i(s,a)]_{i=1}^K$ and matrix $\boldsymbol{\Psi} \in \mathbb{R}^{K \times |\mathcal{S}|}$ whose (s,a)-th column corresponds to $\boldsymbol{\psi}(s,a)^\top$. Further, let $\boldsymbol{P}_{\mathcal{K}} \in \mathbb{R}^{K \times |\mathcal{S}|}$ (resp. $\boldsymbol{\Phi}_{\mathcal{K}} \in \mathbb{R}^{K \times K}$) to be a submatrix of $\boldsymbol{P}$ (resp. $\boldsymbol{\Phi}$) formed by concatenating the rows $\{\boldsymbol{P}_{s,a}, (s,a) \in \mathcal{K}\}$ (resp. $\{\boldsymbol{\Phi}_{s,a}, (s,a) \in \mathcal{K}\}$). By using the previous notations, we can express the relations in Definition 1 and Assumption 1 as $\boldsymbol{P}_{\mathcal{K}} = \boldsymbol{\Phi}_{\mathcal{K}} \boldsymbol{\Psi}$, $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ and $\boldsymbol{\Phi} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\mathcal{K}}$. Note that Assumption 1 suggests $\boldsymbol{\Phi}_{\mathcal{K}}$ is invertible. Taking these equations collectively yields

$$\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi} = \boldsymbol{\Phi}\boldsymbol{\Phi}_{\mathcal{K}}^{-1}\boldsymbol{P}_{\mathcal{K}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\mathcal{K}}\boldsymbol{\Phi}_{\mathcal{K}}^{-1}\boldsymbol{P}_{\mathcal{K}} = \boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}, \tag{14}$$

which is reminiscent of the anchor word condition in topic modelling [2]. In addition, for each iteration t, we denote the collected samples as $\{s_t(s,a)\}_{(s,a)\in\mathcal{K}}$ and define a matrix $\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \in \{0,1\}^{K \times |\mathcal{S}|}$ to be

$$\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}\left((s,a), s'\right) := \begin{cases} 1, & \text{if } s' = s_t(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

for any $(s,a) \in \mathcal{K}$ and $s' \in \mathcal{S}$. Further, we define $\widehat{\boldsymbol{P}}_t = \boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}$. Then it is obvious to see that $\widehat{\boldsymbol{P}}_t$ has nonnegative entries and unit ℓ_1 norm for each row due to Assumption 1, i.e. $\|\widehat{\boldsymbol{P}}_t\|_1 = 1$.

B Analysis of model-based RL (Proof of Theorem 1)

In this section, we will provide complete proof for Theorem 1. As a matter of fact, our proof strategy here justifies a more general version of Theorem 1 that accounts for model misspecification, as stated below.

Theorem 3. *Suppose that $\delta > 0$ and $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Assume that there exists a probability transition model $\widetilde{\boldsymbol{P}}$ obeying Definition 1 and Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \subset \mathbb{R}^K$ and anchor state-action pairs $\mathcal{K}$ such that*

$$\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$$

for some $\xi \ge 0$. Let $\widehat{\pi}$ be the policy returned by Algorithm 1. Assume that

$$N \ge \frac{C \log\left(K/\left((1-\gamma)\delta\right)\right)}{(1-\gamma)^3 \varepsilon^2} \tag{16}$$

for some sufficiently large constant $C > 0$. Then with probability exceeding $1-\delta$,

$$Q^\star(s,a) - Q^{\widehat{\pi}}(s,a) \le \varepsilon + \frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma} + \frac{22\xi}{(1-\gamma)^2}, \tag{17}$$

for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Theorem 3 subsumes Theorem 1 as a special case with $\xi = 0$. The remainder of this section is devoted to proving Theorem 3.

B.1 Proof of Theorem 3

The error $\boldsymbol{Q}^{\widehat{\pi}} - \boldsymbol{Q}^\star$ can be decomposed as

$$\begin{aligned} \boldsymbol{Q}^{\widehat{\pi}} - \boldsymbol{Q}^\star &= \boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}} + \widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star + \widehat{\boldsymbol{Q}}^\star - \boldsymbol{Q}^\star \\ &\ge \boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}} + \widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star + \widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star \\ &\ge -\left(\left\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty + \left\|\widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star\right\|_\infty + \left\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\right\|_\infty\right)\mathbf{1}. \end{aligned} \tag{18}$$

For policy $\widehat{\pi}$ satisfying the condition in Theorem 1, we have $\|\widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star\|_\infty \le \varepsilon_{\mathsf{opt}}$. It boils down to control $\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\|_\infty$ and $\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\|_\infty$.

To begin with, we can use (11) to further decompose $\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\|_\infty$ as

$$\begin{aligned}
\left\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty &= \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \boldsymbol{r} - \left(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}^{\widehat{\pi}}\right)^{-1} \boldsymbol{r}\right\|_\infty \\
&= \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left[\left(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}^{\widehat{\pi}}\right) - \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)\right] \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty \\
&= \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\widehat{\pi}}\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\star}\right\|_\infty + \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^{\star}\right)\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\star}\right|\right\|_\infty + \frac{2\gamma \varepsilon_{\text{opt}}}{1-\gamma}. \qquad (19)
\end{aligned}$$

Here the last inequality is due to

$$\begin{aligned}
&\left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^{\star}\right)\right\|_\infty \\
&\quad \leq \gamma \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\right\|_1 \left\|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^{\star}\right)\right\|_\infty \\
&\quad \leq \gamma \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\right\|_1 \left(\|\boldsymbol{P}\|_1 + \left\|\widehat{\boldsymbol{P}}\right\|_1\right) \left\|\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^{\star}\right\|_\infty \\
&\quad \leq \frac{2\gamma \varepsilon_{\text{opt}}}{1-\gamma},
\end{aligned}$$

where we use the fact that $\|(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}})^{-1}\|_1 \leq 1/(1-\gamma)$ and $\|\boldsymbol{P}\|_1 = \|\widehat{\boldsymbol{P}}\|_1 = 1$.

Similarly, for the term $\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\|_\infty$ in (18), we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\right\|_\infty &= \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^\star}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\pi^\star}\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^\star}\right)^{-1} \left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\pi^\star}\right|\right\|_\infty. \qquad (20)
\end{aligned}$$

As can be seen from (19) and (20), it boils down to bound $|(\boldsymbol{P} - \widehat{\boldsymbol{P}})\widehat{\boldsymbol{V}}^\star|$ and $|(\boldsymbol{P} - \widehat{\boldsymbol{P}})\widehat{\boldsymbol{V}}^{\pi^\star}|$. We have the following lemma.

Lemma 1. *With probability exceeding $1 - \delta$, one has*

$$\begin{aligned}
\left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right)_{s,a} \widehat{\boldsymbol{V}}^\star\right| &\leq \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log(4K/\delta)}{N}} + \frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{(1-\gamma)N} \\
&\quad + \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}} \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)}, \qquad (21) \\
\left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right)_{s,a} \widehat{\boldsymbol{V}}^{\pi^\star}\right| &\leq \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log(4K/\delta)}{N}} + \frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{(1-\gamma)N} \\
&\quad + \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}} \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\pi^\star}\right)}. \qquad (22)
\end{aligned}$$

Proof. See Appendix B.2. □

Applying (21) to (19) reveals that

$$\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N}}\left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)}\right\|_{\infty} + \frac{\gamma}{1-\gamma}\left[4\sqrt{\frac{2\log\left(4K/\delta\right)}{N}}+\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)\,N}\right] + \frac{10\gamma\xi}{(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma}. \tag{23}$$

For the first term, one has

$$\begin{aligned}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\widehat{\pi}}\right)}+\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\star}\right)}\\ &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\left\|\boldsymbol{V}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\\ &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}},\end{aligned}$$

where the first inequality comes from the fact that $\sqrt{\mathsf{Var}(X+Y)} \leq \sqrt{\mathsf{Var}(X)}+\sqrt{\mathsf{Var}(Y)}$ for any random variables X and Y. It follows that

$$\begin{aligned}&\left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)}\right\|_{\infty}\\ &\quad\leq \left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}\right\|_{\infty}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right)\\ &\quad\leq \gamma\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right),\end{aligned} \tag{24}$$

where the second inequality utilizes [3, Lemma 7].

Plugging (24) into (23) yields

$$\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N}}\left[\gamma\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right)\right] + \frac{\gamma}{1-\gamma}\left[4\sqrt{\frac{2\log\left(4K/\delta\right)}{N}}+\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)\,N}\right]+\frac{10\gamma\xi}{(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma}.$$

Then we can rearrange terms to obtain

$$\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq 10\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\left(1-\gamma\right)^3}}+\frac{11\gamma\xi}{(1-\gamma)^2}+\frac{3\gamma\varepsilon_{\text{opt}}}{1-\gamma} \tag{25}$$

as long as $N \geq \widetilde{C}\log(8K/((1-\gamma)\delta))/(1-\gamma)^2$ for some sufficiently large constant $\widetilde{C}>0$.

In a similar vein, we can use (20) and (22) to obtain that

$$\left\|\widehat{\boldsymbol{Q}}^{\pi^{\star}}-\boldsymbol{Q}^{\star}\right\|_{\infty} \leq 10\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\left(1-\gamma\right)^3}}+\frac{11\gamma\xi}{(1-\gamma)^2}. \tag{26}$$

Finally, we can substitute (25) and (26) into (18) to achieve

$$\boldsymbol{Q}^{\widehat{\pi}}-\boldsymbol{Q}^{\star} \geq -\left(20\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\left(1-\gamma\right)^3}}+\frac{22\gamma\xi}{(1-\gamma)^2}+\frac{3\gamma\varepsilon_{\text{opt}}}{1-\gamma}+\varepsilon_{\text{opt}}\right)\mathbf{1}.$$

This result implies that

$$\boldsymbol{Q}^{\widehat{\pi}} \geq \boldsymbol{Q}^{\star}-\left(\varepsilon+\frac{22\xi}{(1-\gamma)^2}+\frac{4\varepsilon_{\text{opt}}}{1-\gamma}\right)\mathbf{1},$$

as long as

$$N \geq \frac{C\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{\left(1-\gamma\right)^3\varepsilon^2},$$

for some sufficiently large constant $C>0$.

B.2 Proof of Lemma 1

To prove this theorem, we invoke the idea of s-absorbing MDP proposed by [1]. For a state $s \in \mathcal{S}$ and a scalar u, we define a new MDP $M_{s,u}$ to be identical to M on all the other states except s; on state s, $M_{s,u}$ is absorbing such that $P_{M_{s,u}}(s|s,a) = 1$ and $r_{M_{s,u}}(s,a) = (1-\gamma)u$ for all $a \in \mathcal{A}$. More formally, we define $P_{M_{u,s}}$ and $r_{M_{u,s}}$ as

$$\begin{aligned} P_{M_{s,u}}(s|s,a) = 1, \quad r_{M_{s,u}}(s,a) = (1-\gamma)u, \qquad & \text{for all } a \in \mathcal{A}, \\ P_{M_{s,u}}(\cdot|s',a') = P(\cdot|s',a'), \quad r_{M_{s,u}}(s,a) = r(s,a), \qquad & \text{for all } s' \neq s \text{ and } a' \in \mathcal{A}. \end{aligned}$$

To streamline notations, we will use $\boldsymbol{V}_{s,u}^{\pi} \in \mathbb{R}^{|\mathcal{S}|}$ and $\boldsymbol{V}_{s,u}^{\star} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the value function of $M_{s,u}$ under policy π and the optimal value function of $M_{s,u}$ respectively. Furthermore, we denote by $\widehat{M}_{s,u}$ the MDP whose probability transition kernel is identical to $\widehat{P}$ at all states except that state s is absorbing. Similar as before, we use $\widehat{\boldsymbol{V}}_{s,u}^{\star} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the optimal value function under $\widehat{M}_{s,u}$. The construction of this collection of auxiliary MDPs will facilitate our analysis by decoupling the statistical dependency between $\widehat{\boldsymbol{P}}$ and $\widehat{\pi}^{\star}$.

To begin with, we can decompose the quantity of interest as

$$\begin{aligned} \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}\right| &= \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}+\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)\right| \\ &\leq \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| + \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)\right| \\ &\overset{\text{(i)}}{\leq} \left|\left(\boldsymbol{P}-\widetilde{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| + \left|\boldsymbol{\lambda}(s,a)\left(\widetilde{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| \\ &\quad + \left|\boldsymbol{\lambda}(s,a)\left(\boldsymbol{P}_{\mathcal{K}}-\widehat{\boldsymbol{P}}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| + \left(\|\boldsymbol{P}_{s,a}\|_1 + \left\|\widehat{\boldsymbol{P}}_{s,a}\right\|_1\right)\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} \\ &\leq \left\|\left(\boldsymbol{P}-\widetilde{\boldsymbol{P}}\right)_{s,a}\right\|_1\left\|\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} + \|\boldsymbol{\lambda}(s,a)\|_1 \cdot \left\|\left(\widetilde{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} \\ &\quad + \|\boldsymbol{\lambda}(s,a)\|_1 \cdot \left\|\left(\boldsymbol{P}_{\mathcal{K}}-\widehat{\boldsymbol{P}}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} + 2\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} \\ &\overset{\text{(ii)}}{\leq} \frac{2\xi}{1-\gamma} + \max_{(s,a)\in\mathcal{K}}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| + 2\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty}, \end{aligned} \tag{27}$$

where (i) makes use of $\widetilde{\boldsymbol{P}}_{s,a} = \boldsymbol{\lambda}(s,a)\widetilde{\boldsymbol{P}}_{\mathcal{K}}$ and $\widehat{\boldsymbol{P}}_{s,a} = \boldsymbol{\lambda}(s,a)\widehat{\boldsymbol{P}}_{\mathcal{K}}$; (ii) depends on $\|\boldsymbol{P}-\widetilde{\boldsymbol{P}}\|_1 \leq \xi$, $\|\boldsymbol{\lambda}(s,a)\|_1 = 1$ and $\|\widehat{\boldsymbol{V}}_{s,u}^{\star}\|_{\infty} \leq (1-\gamma)^{-1}$. For each state s, the value of u will be selected from a set $\mathcal{U}_s$. The choice of $\mathcal{U}_s$ will be specified later. Then for some fixed u in $\mathcal{U}_s$ and fixed state-action pair $(s,a) \in \mathcal{K}$, due to the independence between $\widehat{\boldsymbol{P}}_{s,a}$ and $\widehat{\boldsymbol{V}}_{s,u}^{\star}$, we can apply Bernstein's inequality (cf. [5, Theorem 2.8.4]) conditional on $\widehat{\boldsymbol{V}}_{s,u}^{\star}$ to reveal that with probability greater than $1-\delta/2$,

$$\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| \leq \sqrt{\frac{2\log(4/\delta)}{N}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)} + \frac{2\log(4/\delta)}{3(1-\gamma)N}. \tag{28}$$

Invoking the union bound over all the K state-action pairs of $\mathcal{K}$ and all the possible values of u in $\mathcal{U}_s$ demonstrate that with probability greater than $1-\delta/2$,

$$\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}_{s,u}^{\star}\right| \leq \sqrt{\frac{2\log(4K|\mathcal{U}_s|/\delta)}{N}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)} + \frac{2\log(4K|\mathcal{U}_s|/\delta)}{3(1-\gamma)N}, \tag{29}$$

holds for all state-action pair $(s,a) \in \mathcal{K}$ and all $u \in \mathcal{U}_s$. Here, $\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\cdot)$ is defined in (13). Then we observe that

$$\begin{aligned} \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)} &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right)} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} \\ &\leq \left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}_{s,u}^{\star}\right\|_{\infty} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} \\ &\leq \left|\widehat{\boldsymbol{V}}^{\star}(s)-u\right| + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)}, \end{aligned} \tag{30}$$

where (i) is due to $\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_1+\boldsymbol{V}_2)} \le \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_1)} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_2)}$ and (ii) holds since

$$\left\|\widehat{\boldsymbol{V}}^\star - \widehat{\boldsymbol{V}}^\star_{s,u}\right\|_\infty = \left\|\widehat{\boldsymbol{V}}^\star_{s,\widehat{\boldsymbol{V}}^\star(s)} - \widehat{\boldsymbol{V}}^\star_{s,u}\right\|_\infty \le \left|\widehat{\boldsymbol{V}}^\star(s) - u\right|, \tag{31}$$

whose proof can be found in [1, Lemma 8 and 9].

By substituting (29), (30) and (31) into (27), we arrive at

$$\begin{aligned}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^\star\right| \le{}& \frac{2\xi}{1-\gamma} + \left|\widehat{\boldsymbol{V}}^\star(s)-u\right|\left(2+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\right)\\ &+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)} + \frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{3(1-\gamma)N}.\end{aligned} \tag{32}$$

Then it boils down to determining $\mathcal{U}_s$. The coarse bounds of $\widehat{\boldsymbol{Q}}^{\pi^\star}$ and $\widehat{\boldsymbol{Q}}^\star$ in the following lemma provide a guidance on the choice of $\mathcal{U}_s$.

Lemma 2. *For $\delta \in (0,1)$, with probability exceeding $1-\delta/2$ one has*

$$\left\|\boldsymbol{Q}^\star - \widehat{\boldsymbol{Q}}^{\pi^\star}\right\|_\infty \le \frac{\gamma}{1-\gamma}\sqrt{\frac{\log(4K/\delta)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}, \tag{33}$$

$$\left\|\boldsymbol{Q}^\star - \widehat{\boldsymbol{Q}}^\star\right\|_\infty \le \frac{\gamma}{1-\gamma}\sqrt{\frac{\log(4K/\delta)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}. \tag{34}$$

Proof. See Appendix B.3. □

This inspires us to choose $\mathcal{U}_s$ to be the set consisting of equidistant points in $[\boldsymbol{V}^\star(s)-R(\delta), \boldsymbol{V}^\star(s)+R(\delta)]$ with $|U_s| = \lceil 1/(1-\gamma)^2 \rceil$ and

$$R(\delta) \coloneqq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log(4K/\delta)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}.$$

Since $\|\boldsymbol{V}^\star - \widehat{\boldsymbol{V}}^\star\|_\infty \le \|\boldsymbol{Q}^\star - \widehat{\boldsymbol{Q}}^\star\|_\infty$, Lemma 2 implies that $\widehat{\boldsymbol{V}}^\star(s) \in [\boldsymbol{V}^\star(s)-R(\delta), \boldsymbol{V}^\star(s)+R(\delta)]$ with probability over $1-\delta/2$. Hence, we have

$$\min_{u\in\mathcal{U}_s}\left|\widehat{\boldsymbol{V}}^\star(s)-u\right| \le \frac{2R(\delta)}{|U_s|+1} \le 2\gamma\sqrt{\frac{2\log(4K/\delta)}{N}} + 4\gamma\xi. \tag{35}$$

Consequently, with probability exceeding $1-\delta$, one has

$$\begin{aligned}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^\star\right| &\overset{\text{(i)}}{\le} \frac{2\xi}{1-\gamma} + \min_{u\in\mathcal{U}_s}\left|\widehat{\boldsymbol{V}}^\star(s)-u\right|\left(2+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\right)\\ &\quad+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)} + \frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{3(1-\gamma)N}\\ &\overset{\text{(ii)}}{\le} \frac{2\xi}{1-\gamma} + \left(2\gamma\sqrt{\frac{2\log(4K/\delta)}{N}} + 4\gamma\xi\right)\left(2+\sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\right)\\ &\quad+\sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)} + \frac{2\log\left(8K/\left((1-\gamma)\delta\right)\right)}{3(1-\gamma)N}\\ &\le \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log(4K/\delta)}{N}} + \frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{(1-\gamma)N}\\ &\quad+\sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)},\end{aligned}$$

where (i) follows from (32) and (ii) utilizes (35). This finishes the proof for the first inequality. The second inequality can be proved in a similar way and is omitted here for brevity.

B.3 Proof of Lemma 2

To begin with, one has

$$\begin{aligned}\left\|\left(\widehat{\boldsymbol{P}}-\boldsymbol{P}\right)\boldsymbol{V}^{\star}\right\|_{\infty} &\leq \left\|\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\boldsymbol{\Lambda}\left(\boldsymbol{P}_{\mathcal{K}}-\widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\left(\widetilde{\boldsymbol{P}}-\boldsymbol{P}\right)\boldsymbol{V}^{\star}\right\|_{\infty} \\ &\leq \|\boldsymbol{\Lambda}\|_{1}\left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\|\boldsymbol{\Lambda}\|_{1}\left\|\left(\boldsymbol{P}_{\mathcal{K}}-\widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\widetilde{\boldsymbol{P}}-\boldsymbol{P}\right\|_{1}\|\boldsymbol{V}^{\star}\|_{\infty} \\ &\leq \left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\frac{2\xi}{1-\gamma}, \end{aligned} \tag{36}$$

where the first line uses $\widehat{\boldsymbol{P}}=\boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}$ and $\widetilde{\boldsymbol{P}}=\boldsymbol{\Lambda}\widetilde{\boldsymbol{P}}_{\mathcal{K}}$; the last inequality comes from the facts that $\|\widetilde{\boldsymbol{P}}-\boldsymbol{P}\|_{1}\leq\xi$, $\|\boldsymbol{\Lambda}\|_{1}=1$ and $\|\boldsymbol{V}^{\star}\|_{\infty}\leq(1-\gamma)^{-1}$. Then we turn to bound $\|(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}})\boldsymbol{V}^{\star}\|_{\infty}$. In view of (4), Hoeffding's inequality (cf. [5, Theorem 2.2.6]) implies that for $(s,a)\in\mathcal{K}$,

$$\mathbb{P}\left(\left|\left(\widehat{\boldsymbol{P}}-\boldsymbol{P}\right)_{s,a}\boldsymbol{V}^{\star}\right|\geq t\right)\leq 2\exp\left(-\frac{2t^{2}}{\|\boldsymbol{V}^{\star}\|_{\infty}^{2}/N}\right).$$

Hence by the standard union bound argument we have

$$\left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}\leq\sqrt{\frac{\|\boldsymbol{V}^{\star}\|_{\infty}^{2}\log\left(4K/\delta\right)}{2N}}\leq\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}, \tag{37}$$

with probability over $1-\delta/2$.

1. Now we are ready to bound $\boldsymbol{Q}^{\pi^{\star}}-\widehat{\boldsymbol{Q}}^{\pi^{\star}}$. One has

$$\begin{aligned}\boldsymbol{Q}^{\pi^{\star}}-\widehat{\boldsymbol{Q}}^{\pi^{\star}} &= \left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^{\star}}\right)^{-1}\boldsymbol{r}-\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\boldsymbol{r} \\ &= \left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)-\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^{\star}}\right)\right)\boldsymbol{Q}^{\pi^{\star}} \\ &= \gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}^{\pi^{\star}}-\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)\boldsymbol{Q}^{\pi^{\star}} \\ &= \gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\pi^{\star}}, \end{aligned}$$

   where the first equality makes use of (11). Then we take (36) and (37) collectively to achieve

$$\begin{aligned}\left\|\gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty} &\leq \gamma\sum_{i=0}^{\infty}\left\|\gamma^{i}\left(\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{i}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty} \\ &\leq \gamma\sum_{i=0}^{\infty}\gamma^{i}\left\|\left(\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{i}\right\|_{1}\left\|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty} \\ &\leq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}+\frac{2\gamma\xi}{\left(1-\gamma\right)^{2}}, \end{aligned}$$

   where the last line comes from the fact that for all $i\geq 1$, $(\widehat{\boldsymbol{P}}^{\pi^{\star}})^{i}$ is a probability transition matrix so that $\|(\widehat{\boldsymbol{P}}^{\pi^{\star}})^{i}\|_{1}=1$. This justifies the first inequality (33).

2. In terms of the second one, [1, Section A.4] implies that

$$\left\|\boldsymbol{Q}^{\star}-\widehat{\boldsymbol{Q}}^{\star}\right\|_{\infty}\leq\frac{\gamma}{1-\gamma}\left\|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}.$$

   Substitution of (36) and (37) into the above inequality yields

$$\left\|\boldsymbol{Q}^{\star}-\widehat{\boldsymbol{Q}}^{\star}\right\|_{\infty}\leq\frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}+\frac{2\gamma\xi}{\left(1-\gamma\right)^{2}}.$$

C Analysis of Q-learning (Proof of Theorem 2)

In this section, we will provide complete proof for Theorem 2. We actually prove a more general version of Theorem 2 that takes model misspecification into consideration, as stated below.

Theorem 4. *Consider any $\delta \in (0,1)$ and $\varepsilon \in (0,1]$. Suppose that there exists a probability transition model $\widetilde{\boldsymbol{P}}$ obeying Definition 1 and Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \subset \mathbb{R}^K$ and anchor state-action pairs $\mathcal{K}$ such that*

$$\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$$

for some $\xi \ge 0$. Assume that the initialization obeys $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and for any $0 \le t \le T$, the learning rates satisfy

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}}, \tag{38}$$

for some sufficiently small universal constants $c_1 \ge c_2 > 0$. Suppose that the total number of iterations T exceeds

$$T \ge \frac{C_3 \log(KT/\delta)\log^4 T}{(1-\gamma)^4 \varepsilon^2}, \tag{39}$$

for some sufficiently large universal constant $C_3 > 0$. If there exists a linear probability transition model $\widetilde{\boldsymbol{P}}$ satisfying Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ such that $\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$, then with probability exceeding $1-\delta$, the output Q_T of Algorithm 2 satisfies

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q_T(s,a) - Q^\star(s,a)| \le \varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2}, \tag{40}$$

for some constant $C_4 > 0$. In addition, let π_T (resp. V_T) to be the policy (resp. value function) induced by Q_T, then one has

$$\max_{s\in\mathcal{S}} |V^{\pi_T}(s) - V^\star(s)| \le \frac{2\gamma}{1-\gamma}\left(\varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2}\right). \tag{41}$$

Theorem 4 subsumes Theorem 2 as a special case with $\xi = 0$. The remainder of this section is devoted to proving Theorem 4.

C.1 Proof of Theorem 4

First we show that (41) can be easily obtained from (40). Since [49] gives rise to

$$\|V^{\pi_T} - V^\star\|_\infty \le \frac{2\gamma\|V_T - V^\star\|_\infty}{1-\gamma},$$

we have

$$\|V^{\pi_T} - V^\star\|_\infty \le \frac{2\gamma\|Q_T - Q^\star\|_\infty}{1-\gamma},$$

due to $\|V_T - V^\star\|_\infty \le \|Q_T - Q^\star\|_\infty$. Then (41) follows directly from (40).

Therefore, we are left to justify (40). To start with, we consider the update rule

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1}\right).$$

By defining the error term $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star$, we can decompose $\boldsymbol{\Delta}_t$ into

$$\begin{aligned}
\boldsymbol{\Delta}_t &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1}\right) - \boldsymbol{Q}^\star \\
&= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star) + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\right) \\
&= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star) + \gamma\eta_t\left(\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\right) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) \\
&\quad + \gamma\eta_t(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P})\boldsymbol{V}^\star.
\end{aligned} \tag{42}$$

Here in the penultimate equality, we make use of $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$; and the last equality comes from $\widehat{\boldsymbol{P}}_t = \boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}$ which is defined in (15). It is straightforward to check that $\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}$ is also a probability transition matrix. We denote by $\overline{\boldsymbol{P}} = \boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}$ hereafter. The third term in the decomposition above can be upper and lower bounded by

$$\overline{\boldsymbol{P}}\left(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\right) = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star \le \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}^\star = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1},$$

and

$$\overline{\boldsymbol{P}}\left(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\right) = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star \ge \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star = \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1}.$$

Plugging these bounds into (42) yields

$$\boldsymbol{\Delta}_t \le (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star,$$

$$\boldsymbol{\Delta}_t \ge (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star.$$

Repeatedly invoking these two recursive relations leads to

$$\boldsymbol{\Delta}_t \le \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right), \tag{43}$$

$$\boldsymbol{\Delta}_t \ge \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right), \tag{44}$$

where

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^{t}(1-\eta_j), & \text{if } i = 0, \\ \eta_i\prod_{j=i+1}^{t}(1-\eta_j), & \text{if } 0 < i < t, \\ \eta_t, & \text{if } i = t. \end{cases}$$

Here we adopt the same notations as [4].

To begin with, we consider the upper bound (43). It can be further decomposed as

$$\begin{aligned}\boldsymbol{\Delta}_t \le \underbrace{\eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{(1-\alpha)t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1}\right)}_{=:\boldsymbol{\theta}_t} \\ + \underbrace{\sum_{i=(1-\alpha)t+1}^{t}\eta_i^{(t)}\gamma\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}}_{=:\boldsymbol{\nu}_t} \\ + \underbrace{\sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star}_{=:\boldsymbol{\omega}_t} + \sum_{i=(1-\alpha)t+1}^{t}\eta_i^{(t)}\gamma\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1},\end{aligned} \tag{45}$$

where we define $\alpha := C_4(1-\gamma)/\log T$ for some constant $C_4 > 0$. Next, we turn to bound $\boldsymbol{\theta}_t$ and $\boldsymbol{\nu}_t$ respectively for any t satisfying $\frac{T}{c_2\log\frac{1}{1-\gamma}} \le t \le T$ with stepsize choice (8).

Bounding $\boldsymbol{\omega}_t$. It is straightforward to bound

$$\begin{aligned}\left\|\boldsymbol{\omega}_t\right\|_\infty &\overset{\text{(i)}}{=} \left\|\gamma\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right\|_\infty \\ &\overset{\text{(ii)}}{\le} \gamma\left(\left\|\boldsymbol{\Lambda}\right\|_1\left\|\left(\boldsymbol{P}_{\mathcal{K}} - \widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^\star\right\|_\infty + \left\|\left(\widetilde{\boldsymbol{P}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right\|_\infty\right) \\ &\overset{\text{(iii)}}{\le} \frac{2\gamma\xi}{1-\gamma},\end{aligned}$$

where the first equality comes from the fact that $\sum_{i=1}^{t}\eta_i^{(t)} = 1$ [4, Equation (40)]; the second inequality utilizes $\widetilde{\boldsymbol{P}} = \boldsymbol{\Lambda}\widetilde{\boldsymbol{P}}_{\mathcal{K}}$; the last line uses the facts that $\|\boldsymbol{\Lambda}\|_1 = 1$, $\|\boldsymbol{V}^\star\|_\infty \le (1-\gamma)^{-1}$ and $\|\widetilde{\boldsymbol{P}}_{\mathcal{K}} - \boldsymbol{P}_{\mathcal{K}}\|_1 \le \|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$.

Bounding $\boldsymbol{\theta}_t$. By similar derivation as Step 1 in [4, Appendix A.2], we have

$$\begin{aligned}
\|\boldsymbol{\theta}_t\|_\infty &\le \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{1\le i\le (1-\alpha)t} \eta_i^{(t)} \max_{1\le i\le (1-\alpha)t} \left(\left\| \overline{\boldsymbol{P}}^{\pi_{t-1}} \boldsymbol{\Delta}_{i-1} \right\|_\infty + \left\| \boldsymbol{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \boldsymbol{V}_{i-1} \right\|_\infty + \|\boldsymbol{\Lambda} \boldsymbol{P}_{\mathcal{K}} \boldsymbol{V}_{i-1}\|_\infty \right) \\
&\overset{\text{(i)}}{\le} \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{1\le i\le (1-\alpha)t} \eta_i^{(t)} \max_{1\le i\le (1-\alpha)t} \left(\|\boldsymbol{\Delta}_{i-1}\|_\infty + 2 \|\boldsymbol{V}_{i-1}\|_\infty \right) \\
&\overset{\text{(ii)}}{\le} \frac{1}{2T^2} \cdot \frac{1}{1-\gamma} + \frac{1}{2T^2} \cdot t \cdot \frac{3}{1-\gamma} \\
&\le \frac{2}{(1-\gamma)\, T},
\end{aligned} \tag{46}$$

where (i) is due to the fact that $\|\overline{\boldsymbol{P}}^{\pi_{t-1}}\|_1 = \|\boldsymbol{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}\|_1 = \|\boldsymbol{\Lambda} \boldsymbol{P}_{\mathcal{K}}\|_1 = 1$ and (ii) comes from [4, Equation (39a)].

Bounding $\boldsymbol{\nu}_t$. To control the second term, we apply the following Freedman's inequality.

Lemma 3 (Freedman's Inequality)**.** *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \cdots\}$ with difference sequence $\{X_k : k = 1, 2, 3, \cdots\}$. Assume that the difference sequence is uniformly bounded:*

$$|X_k| \le R \qquad \textit{and} \qquad \mathbb{E}\left[X_k | \{X_j\}_{j=1}^{k-1}\right] = 0 \qquad \textit{for all } k \ge 1.$$

Let

$$S_n := \sum_{k=1}^{n} X_i, \qquad T_n := \sum_{k=1}^{n} \mathsf{Var}\left\{X_k | \{X_j\}_{j=1}^{k-1}\right\}.$$

Then for any given $\sigma^2 \ge 0$, one has

$$\mathbb{P}\left(|S_n| \ge \tau \textit{ and } T_n \le \sigma^2\right) \le 2 \exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right).$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically. For any positive integer $K \ge 1$, with probability at least $1-\delta$ one has

$$|S_n| \le \sqrt{8 \max\left\{T_n, \frac{\sigma^2}{2^K}\right\} \log\frac{2K}{\delta}} + \frac{4}{3} R \log\frac{2K}{\delta}.$$

Proof. See [4, Theorem 4]. □

To apply this inequality, we can express $\boldsymbol{\nu}_t$ as

$$\boldsymbol{\nu}_t := \sum_{i=(1-\alpha)t+1}^{t} \boldsymbol{x}_i,$$

with

$$\boldsymbol{x}_i := \eta_i^{(t)} \gamma \boldsymbol{\Lambda} \left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right) \boldsymbol{V}_{i-1}, \quad \text{and} \quad \mathbb{E}\left[\boldsymbol{x}_i | \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\right] = \boldsymbol{0}. \tag{47}$$

1. In order to calculate bound R in Lemma 3, one has

$$\begin{aligned}
B &:= \max_{(1-\alpha)t<t\le t} \|\boldsymbol{x}_i\|_\infty \le \max_{(1-\alpha)t<t\le t} \left\| \eta_i^{(t)} \boldsymbol{\Lambda} \left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right) \boldsymbol{V}_{i-1} \right\|_\infty \\
&\le \max_{(1-\alpha)t<t\le t} \eta_i^{(t)} \left(\left\| \boldsymbol{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \right\|_1 + \|\boldsymbol{\Lambda} \boldsymbol{P}_{\mathcal{K}}\|_1 \right) \|\boldsymbol{V}_{i-1}\|_\infty \\
&\le \max_{(1-\alpha)t<t\le t} \eta_i^{(t)} \cdot \frac{2}{1-\gamma} \le \frac{4 \log^4 T}{(1-\gamma)^2 T},
\end{aligned}$$

where the last inequality comes from [4, Eqn (39b)] and the fact that $\|\boldsymbol{V}_{i-1}\|_\infty \le \frac{1}{1-\gamma}$.

2. Then regarding the variance term, we claim for the moment that

$$\begin{aligned}\boldsymbol{W}_t &\coloneqq \sum_{i=(1-\alpha)t+1}^{t} \operatorname{diag}\left(\mathsf{Var}\left(\boldsymbol{x}_i|\boldsymbol{V}_{i-1},\cdots,\boldsymbol{V}_0\right)\right)\\ &\le \gamma^2 \sum_{i=(1-\alpha)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_{i-1}\right). \end{aligned} \tag{48}$$

Then we have

$$\begin{aligned}\boldsymbol{W}_t &\le \max_{(1-\alpha)t\le i\le t} \eta_i^{(t)} \left(\sum_{i=(1-\alpha)t+1}^{t} \eta_i^{(t)}\right) \max_{(1-\alpha)t\le i<t} \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right)\\ &\le \frac{2\log^4 T}{(1-\gamma)T} \max_{(1-\alpha)t\le i<t} \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right), \end{aligned} \tag{49}$$

where the second line comes from [4, Eqns (39b), (40)]. A trivial upper bound for $\boldsymbol{W}_t$ is

$$|\boldsymbol{W}_t| \le \frac{2\log^4 T}{(1-\gamma)T} \cdot \frac{1}{(1-\gamma)^2}\mathbf{1} = \frac{2\log^4 T}{(1-\gamma)^3 T}\mathbf{1},$$

which uses the fact that $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \le \|\boldsymbol{V}_i\|_\infty^2 \le 1/(1-\gamma)^2$.

Then, we invoke Lemma 3 with $K = \left\lceil 2\log_2 \frac{1}{1-\gamma}\right\rceil$ and apply the union bound argument over $\mathcal{K}$ to arrive at

$$\begin{aligned}|\boldsymbol{\nu}_t| &\le \sqrt{8\left(\boldsymbol{W}_t + \frac{\sigma^2}{2^K}\mathbf{1}\right)\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}} + \frac{4}{3}B\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}\mathbf{1}\\ &\le \sqrt{8\left(\boldsymbol{W}_t + \frac{2\log^4 T}{(1-\gamma)T}\mathbf{1}\right)\log\frac{8KT}{\delta}} + \frac{4}{3}B\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}\mathbf{1}\\ &\le \sqrt{\frac{32\log^4 T}{(1-\gamma)T}\log\frac{8KT}{\delta}\left(\max_{(1-\alpha)t\le i<t}\mathsf{Var}_{\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}}\left(\boldsymbol{V}_i\right)+\mathbf{1}\right)} + \frac{12\log^4 T}{(1-\gamma)^2 T}\log\frac{8KT}{\delta}\mathbf{1}. \end{aligned} \tag{50}$$

Hence if we define

$$\boldsymbol{\varphi}_t \coloneqq 64\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)T}\left(\max_{\frac{t}{2}\le i\le t}\mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right)+\mathbf{1}\right),$$

then (46) and (50) implies that

$$|\boldsymbol{\theta}_t| + |\boldsymbol{\nu}_t| + |\boldsymbol{\omega}_t| \le \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1}, \tag{51}$$

with probability over $1-\delta$ for all $2t/3 \le k \le t$, as long as $T \gg \log^4 T\log\frac{KT}{\delta}/(1-\gamma)^3$. Therefore, plugging (51) into (45), we arrive at the recursive relationship

$$\boldsymbol{\Delta}_t \le \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1} + \sum_{i=(1-\alpha)k+1}^{k}\eta_i^{(k)}\gamma\overline{\boldsymbol{P}}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} = \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1} + \sum_{i=(1-\alpha)k}^{k-1}\eta_i^{(k)}\gamma\overline{\boldsymbol{P}}^{\pi_{i-1}}\boldsymbol{\Delta}_i.$$

This recursion is expressed in a similar way as [4, Eqn. (46)] so we can invoke similar derivation in [4, Appendix A.2] to obtain that

$$\boldsymbol{\Delta}_t \le 30\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\left(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)}\mathbf{1} + \frac{2\gamma\xi}{(1-\gamma)^2}\mathbf{1}. \tag{52}$$

Then we turn to (44). Applying a similar argument, we can deduce that

$$\boldsymbol{\Delta}_t \ge -30\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\left(1+\max_{\frac{t}{2}\le i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)}\mathbf{1} - \frac{2\gamma\xi}{(1-\gamma)^2}\mathbf{1}. \tag{53}$$

For any t satisfying $\frac{T}{c_2 \log \frac{1}{1-\gamma}} \le t \le T$, taking (52) and (53) collectively gives rise to

$$\|\mathbf{\Delta}_t\|_\infty \le 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}\left(1 + \max_{\frac{t}{2} \le i < t} \|\mathbf{\Delta}_i\|_\infty\right)} + \frac{2\gamma\xi}{(1-\gamma)^2}. \tag{54}$$

Let

$$u_k \coloneqq \max\left\{\|\mathbf{\Delta}_t\|_\infty : 2^k \frac{T}{c_2 \log \frac{1}{1-\gamma}} \le t \le T\right\}.$$

By taking supremum over $t \in \{\lceil 2^k T/(c_2 \log \frac{1}{1-\gamma})\rceil, \ldots, T\}$ on both sides of (54), we have

$$u_k \le 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}(1 + u_{k-1})} + \frac{2\gamma\xi}{(1-\gamma)^2} \qquad \forall\ 1 \le k \le \log\left(c_2 \log \frac{1}{1-\gamma}\right). \tag{55}$$

It is straightforward to bound $u_0 \le \frac{1}{1-\gamma}$. For $k \ge 1$, it is straightforward to obtain from (55) that

$$u_k \le 3\max\left\{30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\}, \tag{56}$$

for $1 \le k \le \log(c_2 \log \frac{1}{1-\gamma})$. We analyze (56) under two different cases:

1. If there exists some integer k_0 with $1 \le k_0 < \lceil \log(c_2 \log \frac{1}{1-\gamma})\rceil$, such that

   $$u_{k_0} \le \max\left\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\right\},$$

   then it is straightforward to check from (56) that

   $$u_{k_0+1} \le 3\max\left\{30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\} \tag{57}$$

   as long as $T \ge C_3(1-\gamma)^{-4}\log^4 T \log(KT/\delta)$ for some sufficiently large constant $C_3 > 0$.

2. Otherwise we have $u_k > \max\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\}$ for all $1 \le k < \lceil \log(c_2 \log \frac{1}{1-\gamma})\rceil$. This together with (56) suggests that

   $$\max\left\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\right\} < 3\max\left\{30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\},$$

   and therefore

   $$\max\left\{30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\} = 30\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}u_{k-1}}$$

   for all $1 \le k \le \log(c_2 \log \frac{1}{1-\gamma})$. Let

   $$v_k \coloneqq 90\sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}u_{k-1}}.$$

   Then we know from (55) that

   $$u_k \le v_k \qquad \forall\ 1 \le k \le \log\left(c_2 \log \frac{1}{1-\gamma}\right).$$

By applying the above two inequalities recursively, we know that

$$\begin{aligned}
u_k \le v_k &= \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2} u_{k-1}^{1/2} \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2} v_{k-1}^{1/2} \\
&\le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2+1/4} u_{k-2}^{1/4} \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2+1/4} v_{k-2}^{1/4} \\
&\le \cdots \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1-1/2^k} u_0^{1/2^k} \le \sqrt{\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\left(\frac{1}{1-\gamma}\right)^{1/2^k},
\end{aligned}$$

where the last inequality holds as long as $T \ge C_3\log^4 T\log(KT/\delta)(1-\gamma)^{-4}$ for some sufficiently large constant $C_3 > 0$. Let $k_0 = \widetilde{c}\log\log\frac{1}{1-\gamma}$ for some properly chosen constant $\widetilde{c} > 0$ such that k_0 is an integer between 1 and $\log(c_2\log\frac{1}{1-\gamma})$, we have

$$u_{k_0} \le \sqrt{\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\left(\frac{1}{1-\gamma}\right)^{1/2^{k_0}} = O\left(\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\right).$$

When $T \ge C_3\log^4 T\log(KT/\delta)(1-\gamma)^{-4}$ for some sufficiently large constant $C_3 > 0$, this implies that $u_{k_0} < 1$, which contradicts with the preassumption that $u_k > \max\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\}$ for all $1 \le k \le c_2\log\frac{1}{1-\gamma}$.

Consequently, (57) must hold true and then the definition of u_k immediately leads to

$$\|\boldsymbol{\Delta}_T\|_\infty \le 90\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}} + \frac{6\gamma\xi}{(1-\gamma)^2}.$$

Then for any $\varepsilon \in (0, 1]$, one has

$$\|\boldsymbol{\Delta}_T\|_\infty \le \varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2},$$

as long as

$$90\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}} \le \varepsilon.$$

Hence, if the total number of iterations T satisfies

$$T \ge C_3\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4\varepsilon^2}$$

for some sufficiently large constant $C_3 > 0$, (10) would hold for Algorithm 1 with probability over $1-\delta$.

Finally, we are left to justify (48). Recall the definition of $\boldsymbol{x}_i$ (cf. (47)), one has

$$\begin{aligned}
\mathrm{diag}\left(\mathsf{Var}\left(\boldsymbol{x}_i|\boldsymbol{V}_{i-1},\cdots,\boldsymbol{V}_0\right)\right) &= \gamma^2\left(\eta_i^{(t)}\right)^2\mathrm{diag}\left(\mathsf{Var}\left(\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}|\boldsymbol{V}_{i-1}\right)\right) \\
&= \gamma^2\left(\eta_i^{(t)}\right)^2\mathrm{diag}\left(\boldsymbol{\Lambda}\mathsf{Var}\left(\left(\hat{\boldsymbol{P}}_{\mathcal{K}}^{(i)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}|\boldsymbol{V}_{i-1}\right)\boldsymbol{\Lambda}^\top\right) \\
&= \gamma^2\left(\eta_i^{(t)}\right)^2\left\{\boldsymbol{\lambda}(s,a)^2\,\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}\left(\boldsymbol{V}_{i-1}\right)\right\}_{s,a},
\end{aligned}$$

where the notation $\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}(\boldsymbol{V}_{i-1})$ is defined in (12). Plugging this into the definition of $\boldsymbol{W}_t$ leads to

$$\begin{aligned}
\boldsymbol{W}_t &= \gamma^2\sum_{i=(1-\alpha)t+1}^{t}\left(\eta_i^{(t)}\right)^2\left\{\boldsymbol{\lambda}(s,a)^2\,\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}\left(\boldsymbol{V}_{i-1}\right)\right\}_{s,a} \\
&= \gamma^2\sum_{i=(1-\alpha)t+1}^{t}\left(\eta_i^{(t)}\right)^2\left\{\boldsymbol{\lambda}(s,a)^2\left(\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V}_{i-1}\circ\boldsymbol{V}_{i-1}\right) - \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\circ\left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\right)\right\}_{s,a}. \quad (58)
\end{aligned}$$

Then we introduce a useful claim as follows. The proof is deferred to Appendix C.2.

Claim 1. For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, one has

$$\begin{aligned}
&\boldsymbol{\lambda}(s,a)^2 \left(\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \circ (\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\right) \\
&\quad \leq \boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \circ (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}). \qquad (59)
\end{aligned}$$

By invoking this claim with $\boldsymbol{V} = \boldsymbol{V}^{i-1}$ and taking collectively with (58), one has

$$\begin{aligned}
\boldsymbol{W}_t &\leq \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \left\{\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}_{i-1} \circ \boldsymbol{V}_{i-1}) - (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}) \circ (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1})\right\}_{s,a} \\
&= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \left[\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}_{i-1} \circ \boldsymbol{V}_{i-1}) - (\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}) \circ (\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1})\right] \\
&= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \mathsf{Var}_{\overline{\boldsymbol{P}}}(\boldsymbol{V}_{i-1}),
\end{aligned}$$

which is the desired result.

C.2 Proof of Claim 1

To simplify notations in this proof, we use $[\lambda_i]_{i=1}^K$, $[P_{i,j}]_{1\leq i\leq K, 1\leq j\leq |\mathcal{S}|}$ and $[V_i]_{i=1}^{|\mathcal{S}|}$ to denote $\boldsymbol{\lambda}(s,a)$, $\boldsymbol{P}_{\mathcal{K}}$ and $\boldsymbol{V}$ respectively. Then one has

$$\begin{aligned}
&\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \circ (\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \\
&\quad - \boldsymbol{\lambda}(s,a)^2 \left(\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V} \circ \boldsymbol{V}) - (\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \circ (\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\right) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j^2 - \left(\sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j\right)^2 - \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i^2 P_{i,j} V_j^2 + \sum_{i=1}^{K} \lambda_i^2 \left(\sum_{j=1}^{|\mathcal{S}|} P_{i,j} V_j\right)^2 \\
&= \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \left[(1-\lambda_i) V_j - \sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} V_{j'}\right]. \\
&= \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \left[\left(\sum_{i'=1}^{K}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} - \lambda_i\right) V_j - \sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} V_{j'}\right] \\
&= \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|}\sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \lambda_{i'} P_{i',j'} (V_j - V_{j'})
\end{aligned}$$

where in the penultimate equality, we use the fact that

$$\sum_{i'=1}^{K}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} = \boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{1} = 1.$$

It follows that

$$
\begin{aligned}
&\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\circ(\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}) \\
&\quad - \boldsymbol{\lambda}(s,a)^2\left(\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\circ(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\right) \\
&= \sum_{i=1}^{K}\sum_{1\le i'<i}\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}\left[\lambda_i P_{i,j}V_j\lambda_{i'}P_{i',j'}(V_j - V_{j'}) + \lambda_{i'}P_{i',j}V_j\lambda_i P_{i,j'}(V_j - V_{j'})\right] \\
&= \sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}V_jP_{i',j'}(V_j - V_{j'}) + \sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i',j}V_jP_{i,j'}(V_j - V_{j'})\right] \\
&\overset{\text{(i)}}{=} \sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}V_jP_{i',j'}(V_j - V_{j'}) + \sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i',j'}V_{j'}P_{i,j}(V_{j'} - V_j)\right] \\
&= \sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}P_{i',j'}(V_j - V_{j'})^2\right] \\
&\ge 0,
\end{aligned}
$$

where in (i), we exchange the indices j and j'.

D Feature dimension and the number of anchor state-action pairs

The assumption that the feature dimension (denoted by K_{d}) and the number of anchor state-action pairs (denoted by K_{n}) are equal is actually non-essential. In what follows, we will show that if $K_{\mathsf{d}} \neq K_{\mathsf{n}}$, then we can modify the current feature mapping $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{K_{\mathsf{d}}}$ to achieve a new feature mapping $\phi' : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{K_{\mathsf{n}}}$ that does not change the transition model P. By doing so, the new feature dimension K_{n} equals to the number of anchor state-action pairs.

To begin with, we recall from Definition 1 that there exists K_{d} unknown functions $\psi_1, \cdots, \psi_{K_{\mathsf{d}}} : \mathcal{S} \to \mathbb{R}$, such that

$$
P(s'|s,a) = \sum_{k=1}^{K_{\mathsf{d}}}\phi_k(s,a)\,\psi_k(s'),
$$

for every $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \in \mathcal{S}$. In addition, we also recall from Assumption 1 that there exists $\mathcal{K} \subseteq \mathcal{S}\times\mathcal{A}$ with $|\mathcal{K}| = K_{\mathsf{n}}$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$
\phi(s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}}\lambda_i(s,a)\,\phi(s_i,a_i) \in \mathbb{R}^{K_{\mathsf{d}}} \quad \text{for} \quad \sum_{i=1}^{K_{\mathsf{n}}}\lambda_i(s,a) = 1 \quad \text{and} \quad \lambda_i(s,a) \ge 0.
$$

Case 1: $K_{\mathsf{d}} > K_{\mathsf{n}}$. In this case, the vectors in $\{\phi(s,a) : (s,a) \in \mathcal{K}\}$ are linearly independent. For ease of presentation and without loss of generality, we assume that $K_{\mathsf{d}} = K_{\mathsf{n}} + 1$. This indicates that the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{K_{\mathsf{d}}\times(|\mathcal{S}||\mathcal{A}|)}$ whose columns are composed of the feature vectors of all state-action pairs has rank K_{n} and is hence not full row rank. This suggests that there exists K_{n} linearly independent rows (without loss of generality, we assume they are the first K_{n} rows). We can remove the last row from $\boldsymbol{\Phi}$ to obtain $\boldsymbol{\Phi}' := \boldsymbol{\Phi}_{1:K_{\mathsf{n}},:} \in \mathbb{R}^{K_{\mathsf{n}}\times(|\mathcal{S}||\mathcal{A}|)}$ such that $\boldsymbol{\Phi}'$ is full row rank. Then we show that we can actually use the columns of $\boldsymbol{\Phi}'$ as new feature mappings. To see why this is true, note that the last row $\Phi_{K_{\mathsf{n}}+1,:}$ can be represented as a linear combination of the first K_{n} rows, namely there must exist constants $\{c_k\}_{k=1}^{K_{\mathsf{n}}}$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$
\phi_{K_{\mathsf{n}}+1}(s,a) = \sum_{k=1}^{K_{\mathsf{n}}} c_k\phi_k(s,a).
$$

Define $\psi'_k = \psi_k + c_k \psi_{K_{\mathsf{n}}+1}$ for $k = 1, \ldots, K_{\mathsf{n}}$, we have

$$P(s'|s,a) = \sum_{k=1}^{K_{\mathsf{d}}} \phi_k(s,a)\,\psi_k(s') = \phi_{K_{\mathsf{n}}+1}(s,a)\,\psi_{K_{\mathsf{n}}+1}(s') + \sum_{k=1}^{K_{\mathsf{n}}} \phi_k(s,a)\,\psi_k(s')$$
$$= \sum_{k=1}^{K_{\mathsf{n}}} \phi_k(s,a)\,[\psi_k(s') + c_k \psi_{K_{\mathsf{n}}+1}(s')] = \sum_{k=1}^{K_{\mathsf{n}}} \phi_k(s,a)\,\psi'_k(s'),$$

which is linear with respect to the new K_{n} dimensional feature vectors. It is also straightforward to check that the new feature mapping satisfies Assumption 1 with the original anchor state-action pairs $\mathcal{K}$.

Case 2: $K_{\mathsf{d}} < K_{\mathsf{n}}$. For ease of presentation and without loss of generality, we assume that $K_{\mathsf{n}} = K_{\mathsf{d}} + 1$ and that the subspace spanned by the feature vectors of anchor state-action pairs is non-degenerate, i.e., has rank K_{d} (otherwise we can use similar method as in Case 1 to further reduce the feature dimension K_{d}). In this case, the matrix $\mathbf{\Phi}_{\mathcal{K}} \in \mathbb{R}^{K_{\mathsf{d}} \times K_{\mathsf{n}}}$ whose columns are composed of the feature vectors of anchor state-action pairs has rank K_{d}. We can add $K_{\mathsf{n}} - K_{\mathsf{d}} = 1$ new row to $\mathbf{\Phi}_{\mathcal{K}}$ to obtain $\mathbf{\Phi}'_{\mathcal{K}} \in \mathbb{R}^{K_{\mathsf{n}} \times K_{\mathsf{n}}}$ such that $\mathbf{\Phi}'_{\mathcal{K}}$ has full rank K_{n}. Then we let the columns of $\mathbf{\Phi}'_{\mathcal{K}} = [\phi'(s,a)]_{(s,a)\in\mathcal{K}}$ to be the new feature vectors of the anchor state-action pairs, and define the new feature vectors for all other state-action pairs $(s,a) \notin \mathcal{K}$ by

$$\phi'(s,a) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i(s,a)\,\phi'(s_i,a_i).$$

We can check that the transition model P is not changed if we let $\psi_{K_{\mathsf{n}}}(s') = 0$ for every $s' \in \mathcal{S}$. It is also straightforward to check that Assumption 1 is satisfied.

To conclude, when $K_{\mathsf{d}} \neq K_{\mathsf{n}}$, we can always construct a new set of feature mappings with dimension K_{n} such that: (i) the feature dimension equals to the number of anchor state-action pairs (they are both K_{n}); (ii) the transition model can still be linearly parameterized by this new set of feature mappings; and (iii) the anchor state-action pair assumption (Assumption 1) is satisfied with the original anchor state-action pairs.

References

[1] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[2] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[4] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[5] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.